DOCUMENT RESUME

ED 146 192 TM 006 562

AUTHOR Gillmore, Gerald M.; And Others
TITLE The Generalizability of Student Instructional Ratings: General Theory and Application to the University of Washington Instructional Assessment System.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
REPORT NO EAC-76-14
PUB DATE May 76
NOTE 32p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Analysis of Variance; College Students; College Teachers; *Course Evaluation; Data Collection; *Error Patterns; Higher Education; Sampling; *Statistical Analysis; *Student Evaluation of Teacher Performance; *Test Reliability
IDENTIFIERS *Generalizability Theory; University of Washington Instructional Assessment

ABSTRACT

Generalizability theory is described and applied to data collection methods for student ratings of instruction. Three generalizability coefficients (generalizing over both students and items, generalizing over students only, and generalizing over items only) are defined, and the appropriateness of each for various purposes is discussed. Generalizability theory is applied to data derived from the University of Washington Instructional Assessment System. Results show that adequate dependability for evaluating teacher or course effectiveness is obtained when fifteen or more students rate a class on four or five items. A single item is reliable for diagnosing instructional problems with 15 or more student raters. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED146192

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 562

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C.E. Lunneborg

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Educational Assessment Center

University of Washington

May 1976

76-14

The Generalizability of Student Instructional Ratings: General Theory and Application to the University of Washington Instructional Assessment System

Gerald M. Gillmore, Michael T. Kane* & Richard W. Naccarato

Generalizability theory is described and applied to the design under which data are collected for student ratings of instruction. Specifically, we defined three generalizability coefficients and discussed the appropriateness of each for various uses of ratings data. The theory is then applied to data deriving from the UW Instructional Assessment System. Results show that adequate dependability for evaluating teacher/course effectiveness is obtained if fifteen or more students rate a class with four or five items. A single item has adequate dependability for diagnosing instructional problems with fifteen or more student raters.

Educational Assessment Center Project: 506

*Michael Kane is at SUNY at Stony Brook.

The Generalizability of Student Instructional Ratings: General Theory and Application to the University of Washington Instructional Assessment System

Gerald M. Gillmore, Michael T. Kane and Richard W. Naccarato

In classical reliability theory (Gulliksen, 1950) an observed score is assumed to have two components. One of these components, the true score (T) represents the actual value of the attribute being measured. The other component is random error (E). If we call an observed score X, then $X_i = T_i + E_i$ where i represents the i^{th} subject. A further assumption is that true score and error are uncorrelated for a given test and, thus, the total variance of the observed scores can be partitioned into true score variance and error variance, i.e. $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. The reliability of a test is then defined as the ratio σ_T^2 / σ_X^2, or $1 - (\sigma_E^2 / \sigma_X^2)$. It can be equivalently defined as the squared correlation between the observed scores and the true scores.

Generalizability theory (Cronbach et al, 1972) is a broadening of classical reliability theory to take advantage of information which can be provided by data collected with a more complex design. Whereas the observed score variance is partitioned into only two sources in all cases within reliability theory, it can be partitioned into multiple sources within generalizability theory, depending upon the design under which the data were collected.

Suppose, for example, 100 subjects wrote stories in response to each of four Thematic Apperception Test pictures. Furthermore, suppose four raters rated all stories for achievement motivation. In analysis of variance terminology, our example is a three-way completely crossed design; subjects crossed with raters crossed with pictures. One can compute separate sums of squares, mean squares, and most importantly here, variance components for three main effects, three two-way interactions, and one three-way interaction. A design

such as this causes real problems for classical reliability theory in terms of which of these variances is the true score variance and which is the error variance. Clearly one could compute an inter-rater reliability or an inter-picture reliability and not expect the two to be equal in magnitude and certainly realize that the two are not equivalent. The reason for the non-equivalence is that true score assumes a different definition for each of the two reliabilities, as does error. Furthermore, a third distinct reliability coefficient which treats the variance components of pictures, raters, and all interactions as error is possible. Classical reliability theory could treat this possibility but would not naturally do so.

Thus, while classical reliability theory can be made to handle most situations, it becomes cumbersome and confusing when variation can be partitioned into more than two sources. Misinterpretations are likely, and there is little chance to gain information which will be helpful in the designing of future studies for increased efficiency.

Generalizability theory, in contrast, explicitly adopts a multifaceted analysis of error as a standard technique. Such analyses are not new to generalizability theory, but were previously presented by Lindquist (1953). To illustrate what is meant by treating error as multifacted, in our above example there are potentially six sources of error: (1) the main effect of raters, (2) the main effect of pictures, and (3) - (6) all interactions. The only source of variance which would not be considered error in any reasonable context is the subject variance, since subject differences are that which the measuring device is seeking to detect.

Cronbach, et al, (1972) listed the following advantages of a multifaceted analysis of error:

> 1. Explicit consideration of the several facets of a measuring operation dispels ambiguities that were present in, and concealed by, the classical models.
>
> 2. The multifacet study can appraise interactions inaccessible to the older methods, and so can improve one's understanding of the measure.
>
> 3. One multifacet study answers questions that formerly required several separate sets of data.
>
> 4. Multifacet information enables one to design more efficient procedures for collecting data, either for the measurement of individuals or for the determination of group means. (p 2).

There are several additional ways in which generalizability theory departs from classical reliability theory. For one, the "true" score of the latter is replaced by the "universe" score. The universe score is defined as the mean of all admissable measurements which could be made on an individual. Since in practice it is impossible to obtain all admissable measurements the universe score is estimated by the mean of all observations in a sample. In using the observed score to estimate the universe score, the investigator is attempting to generalize from a sample of observations to a universe. Reliability considerations then become questions of precision of generalization, and hence the name generalizability.

Universe scores replace the true score of classical reliability theory and are preferred because they place emphasis on both the presence of many universes to which one can generalize, and the importance of specifying to which universe it is. True score, on the other hand, has a strong unitary connotation and, as Cronbach et al (1947) noted, "... no one thinks to inquire 'Which truth?' " (p 19).

Generalizability theory also emphasizes a distinction between G and D studies. Studies carried out in order to determine the generalizability of the observations are called G studies, while studies upon which decisions will ultimately be made are called D studies. One study can serve both functions,

but when a G study is conducted in advance, a D study can be designed to be more efficient. The G study, D study distinction is partially an extension of the Spearman-Brown prophecy formula, which is most commonly used to assess the effect of lengthening or shortening a test on inter-item reliability. To revert again to our earlier example, recall that raters were crossed with subjects. It might be determined from the G study that adequate generalizability could be obtained for a D study by having subjects nested within raters, eg, each rater rates one-fourth of the subjects. It might also be determined that only 2 pictures are needed, or maybe 8 pictures.

What we earlier loosely called "all admissable measurements" is more formally defined as the <u>universe of admissable observations</u> by Cronbach et al (1972). The universe score is the mean over this universe. It is useful to think of the universe as a multivariate space of observations. Each dimension is referred to as a facet and each facet can take on a number of possible conditions. In our prior example, we had two facets. In order to specify an acceptable observation for a subject, a particular condition would have to be specified for the rater facet and a particular condition would have to be specified for the picture facet. To get ahead of ourselves a bit, student evaluations of teachers are also characterized by at least two facets, an item facet and a student facet. We define an admissable observation on a given teacher by specifying a particular student and a particular item. In general, each facet is defined as a set of conditions and the universe of admissable observations is defined as the set of all possible combinations of conditions, taking one condition from each facet.

Coefficients of generalizability, $\varepsilon\rho^2$, are defined as the ratio of universe score variance to the expected observed score variance, and are approximately

equal to the squared correlation of observed score and universe score (Cronbach, et al, 1972, p 82). These coefficients are essentially intraclass correlation coefficients (Cronbach, 1964).

The Application of Generalizability Theory to Student Course Evaluations

In the standard paradigm for student course evaluations, a common set of evaluative items is administered to the students within each class being evaluated. The purpose of the instrument as a whole, and each item within it is to provide evaluative information about the course and the instructor. Hence, it is the dependability of class means rather than individual student responses that is of interest, and the class is thereby the appropriate unit of analysis.

This design is such that each student responds to the same items, regardless of which class is being rated. However, a different set of students rate each class. Thus, students are nested within classes and crossed with items. This design is often labelled split-plot, and is designated as V-B by Cronbach et al, (1972, p. 38 and pp. 189-193). Note, however, that the unit of analysis, designated as p (for persons) by Cronbach et al is designated as c for classes in our analyses.

The random effects analysis of variance for design V-B, with students (s) nested within classes (c), and both classes and students crossed with items (i), yields estimates of five components of variance. Expected mean square formulas from which these variance component estimates can be computed are presented in Table 1.

Table I here

The layout in Table 1 assumes a balanced analysis of variance (i.e. the same number of students is sampled from each class, and each student responds to all items) and with all effects being random (i.e. all effects are sampled from infinite universes). The constants j, k, and n are, respectively, the numbers of classes, items, and students per class which are sampled to estimate the variance components.

All of the components of variance to be presented in this paper are derived from the random effects design presented in Table 1. It is possible to base the analysis of a G-study on a fixed effects model or a mixed model with equivalent results (Kane and Brennan, in press), but consideration of these alternatives would complicate this discussion unnecessarily.

Once the universe of admissable observations is defined and the variance components are estimated in the G study, one must focus on the purpose of the D-study in order to define a generalizability coefficient. The design of the D-study then has two important implications. First, the D-study determines which components of variance are included in the observed score variance estimate. In some cases, the observed score variance estimate will exhaust the set of variance components; however, such is not the case here. Recall that the students within each class responds to the same set of items for the G-study. It is our contention that this will also be true for almost all applications. In such a case, the mean for each class is computed over the same set of items and the item main effect makes no contribution to the observed score variance (see Ebel, 1965). If, for a D-study, a different set of items were to be randomly selected for each class, then the variance component from the items main effect would have to be included. (If items were nested within classes in the G-study, the class by item interaction would be confounded with the item main effect, and the investigator would be divested of his opportunity to <u>not</u> include

the latter in the observed score variance.)

Assuming all D-studies will use a common set of items for all classes, the observed score variance, which is also the denominator for all generalizability coefficients, is the following:

$$\sigma^2(X) = \sigma^2(c) + \frac{1}{k'}\,\sigma^2(ci) + \frac{1}{n'}\,\sigma^2(s, cs) + \frac{1}{n'k'}\,\sigma^2(e) \quad (1)$$

Formula 1 follows the notation of Table 1 except k' and n' refer to the number of items and students, respectively, which are employed in the D-study. If an estimate of the observed score for the G-study is needed, formula 1 can be used with k' = k and n' = n.

In formula 1, $\sigma^2(c)$ is not divided by the number of classes since classes are the units of analysis and hence, the variance component of classes is not dependent in any systematic way upon the number of classes sampled. However, the remaining variance components are systematically reduced in magnitude as the number of conditions of each is increased within a class and, hence, $\sigma^2(ci)$, $\sigma^2(s,cs)$ and $\sigma^2(e)$ are divided by k', n' and the product of the two respectively.

The second implication of the D-study design is the universe to which the investigator wishes to generalize. Definition of the universe score is necessary for determining the proper partitioning of the observed score variance into universe score variance and error variance. If the conditions of a facet in a study exhaust the set of conditions defining the facet, then the universe score is the average over this set of conditions, and there is no intention to generalize to a larger set of conditions for that facet. If the conditions are taken to be a sample from an infinite set, then the universe score is the average over this larger set and generalization is to the larger set. For course evaluations, the universe of admissible observations has two facets, and for

each facet, the conditions in the D-study can be taken as a sample from some larger set of conditions, or can be said to define the full set of conditions for the facet. Thus, there are four possibilities, and each of these possibilities leads to a different definition of universe score and hence to a different generalizability coefficient. However, the coefficient that arises when the investigator chooses to generalize over neither items nor students is not estimable from the data yielded by design V-B (see Kane & Brennan, in press).

Before discussing the three remaining coefficients, a caveat should be introduced. In using course evaluation questionnaires, the focus is often on the quality of the instructor. In the data analyzed here, the instructor effect is confounded with the effect of the course content, with the effect of differential self-selection of students into classes, with the effect of occasion (our data were collected during the last two weeks of the semester, in a class period chosen by the instructor), and with any other special effects associated with a particular section. All of these effects will inflate the variance attributed to instructor to an unknown extent, and this will tend to inflate all of the coefficients. It is, therefore, more precise to talk about the generalizability of the results for an instructional event, or class, rather than for an instructor, and we refer to the class effect, rather than the instructor effect.

Generalizability Estimate 1--Generalizing Over Both Students and Items

In generalizing over both students and items, the universe score for any class is the expected value of the average rating over all possible samples of admissible items and admissible students. It is assumed that the items used are drawn randomly from an infinite set of admissible items, and that the students responding to the items are drawn randomly from an infinite set of admissible students.

Since the universe score is the expected value of the observed score, taken over the universe of admissible observations, the universe score variance may be obtained for this case by taking the limit of the observed score variance

as both k' and n' go to infinity. Generalizing over both students and items, the appropriate generalizability coefficient is:

$$\varepsilon\rho^2(S,I) = \frac{\sigma^2(c)}{\sigma^2(c) + \frac{1}{k'}\sigma^2(ci) + \frac{1}{n'}\sigma^2(s,cs) + \frac{1}{n'k'}\sigma^2(e)} \qquad (2)$$

More detailed derivations of the coefficients discussed in this paper can be found in Gleser et al. (1965), Lindquist (1953), and Kane and Brennan (in press).

In formula (1), the universe score variance is the variance of the main effect due to classes, $\sigma^2(c)$. All of the other components that enter the observed score variance are taken as sources of error. $\sigma^2(c)$ corresponds to true score variance in classical test theory, and the remaining portion of the observed score variance corresponds to error variance.

In classical test theory, the error variance is undifferentiated, and increasing the number of observations that are averaged to obtain observed scores by a factor of, say, m, decreases the error variance by its reciprocal, $\frac{1}{m}$. This regularity is the basis for the Spearman-Brown formula for changes in the length of a test. In Formula (1), the error variance has no such simple relationship to the number of students, items, or the product of the two, and the Spearman-Brown formula does not apply. It is, however, easy to estimate the value of $\varepsilon\rho^2(S, I)$ for any number of students and items by substituting appropriate values for n' and k' in Formula (2).

$\varepsilon\rho^2(S,I)$ is approximately equal to the expected value of the correlation of two evaluations based on independent random samples of k' items and independent random samples of n' students. The sampling variances due to both students and items are included in the error variance.

Generalizability Estimate 2: Generalizing over Students

A second generalizability coefficient results if we generalize over students

but not over items. In this case, the universe score for a class is the expected value, over samples of students, of the class average rating on the set of items used in the D-study. The measure of interest is the instructor's rating on a particular set of items, and we do not consider the items used in the D-study to be a sample from any larger set of items. The items used in the D-study define a restricted universe of conditions for the item effect.

The decision to generalize to this restricted universe does not change the expected observed score variance, but does change the universe score variance. The universe score is the expected value of the observed score over an infinite universe of students and a finite universe of k' items. The universe score variance is found by taking the limit of the observed score variance as the number of students, n', goes to infinity. The generalizability coefficient, for generalization over students but not over items, is given by:

$$\varepsilon\rho^2(S) = \frac{\sigma^2(c) + \frac{1}{k'}\sigma^2(ci)}{\sigma^2(c) + \frac{1}{k'}\sigma^2(ci) + \frac{1}{n'}\sigma^2(s,cs) + \frac{1}{n'k'}\sigma^2(e)} \qquad (3)$$

where k' and n' are the numbers of items and students per class in the D-study, and the variance components are estimated by a random effects analysis of the G-study.

In the previous case, where the universe score was defined as the expected value over an infinite set of items, the variance attributable to the sampling of the class-item interaction was taken to be error. In this case, the items in the D-study define the universe of items, and there can be no error due to the sampling of class-item interactions. The variance, $\frac{1}{k'}\sigma^2(ci)$, that was considered error in formula (2) is now part of the universe score variance. The differential responses of different classes to the D-study items is a legitimate part of the universe score.

Note that while the variance components in formula (3) are derived from a G-study analysis that assumes that items are sampled from an infinite universe of items, it is still possible to estimate the appropriate generalizability coefficient for a D-study in which the items are assumed to exhaust the universe of items.

In Formula (3) $\left[\sigma^2(c) + \frac{1}{k'}\sigma^2(ci)\right]$ corresponds to the true score variance of classical test theory, and $\frac{1}{n'}\left[\sigma^2(s,cs) + \frac{1}{k'}\sigma^2(e)\right]$ corresponds to the error variance. Increasing the number of students upon which the class average is based by a factor m will decrease the error variance by $\frac{1}{m}$ and the Spearman-Brown formula holds for the number of students. Increasing the number of items by m, however, does <u>not</u> decrease the error variance by $\frac{1}{m}$ and the Spearman-Brown does not apply for items.

$\varepsilon\rho^2(s)$ is approximately equal to the expected value of the correlation of any two measures of class quality based on the same k' items and independent samples of n' students each. It is an estimate of average inter-rater agreement.

<u>Generalizability Estimate 3: Generalizing over Items</u>

A third generalizability coefficient is obtained if we generalize over items, but not over students. The universe score is the expected value of the average rating given by a fixed set of students to a random sample of items. The students used in the D-study are not considered to be a sample from a larger universe of students.

Restricting interest to a particular group of students for each instructor, and generalizing over items, the generalizability coefficient is

$$\varepsilon\rho^2(I) = \frac{\sigma^2(c) + \frac{1}{n'}\sigma^2(s,cs)}{\sigma^2(c) + \frac{1}{n'}\sigma^2(s,cs) + \frac{1}{k'}\sigma^2(ci) + \frac{1}{n'k'}\sigma^2(e)}$$

Since the student-class variance and the student variance contribute to both

the universe score variance and the observed score variance, the confounding of the student variance with the class-student variance in the G-study does not interfere with the estimation of $\varepsilon\rho^2(I)$. In this case, the Spearman-Brown formula is valid for changes in the number of items but not for changes in the number of students.

$\varepsilon\rho^2(I)$ is approximately equal to the expected correlation between any two measures of teacher performance based on a common sample of students but on independent ramdom samples of items. It is a measure of the internal consistency of the items.

Alternative Formulas

Earlier we stated that generalizability coefficients are in essence, intra-class correlations. The equivalence becomes more evident if we present formulas for the three generalizability coefficients introduced above in terms of mean squares. Importantly, these formulas only hold in the case where $n' = n$ and $k' = k$.

$$\varepsilon\rho^2(S,I) = \frac{MS(c) - MS(s, cs) - MS(c,i) + MS(e)}{MS(c)} \qquad (5)$$

$$\varepsilon\rho^2(S) = \frac{MS(c) - MS(s,cs)}{MS(c)} \qquad (6)$$

$$\varepsilon\rho^2(I) = \frac{MS(c) - MS(ci)}{MS(c)} \qquad (7)$$

Comparison of the Three Coefficients

Each of the three generalizability coefficients is potentially a legitimate estimate of the dependability of student course evaluations. The specific one which should be used depends upon the purpose of the D-study. In attempting a general assessment of teaching effectiveness as perceived by students, $\varepsilon\rho^2(S,I)$

would seem to be the most appropriate coefficient. Its use is indicated when generalization over both items and students is desired. Generalizing over students is certainly appropriate since the evaluation given a class by a particular group of students is not of primary interest. The purpose in evaluating teachers or curricula is to make decisions which go beyond that one particular offering, and seldom do the same students take a given course more than once. In order to interpret the results of a course evaluation procedure realistically, it is necessary to know how dependably the average rating given by a sample of students estimates the average rating that would be given by all of the students who might reasonably have taken the course (and may, in fact, take it in the future). The desirability of generalizing over both items and students, rather than students alone, is less obvious since it is likely that the same set of items, on a printed form, would be used for all classes in an institution over several years. However, our interest is seldom confined to a particular set of items in assessing teaching effectiveness, but to all possible items which could have been used.

While we favor use of $\varepsilon\rho^2(S,I)$ for assessing the dependability of rating for general instructional evaluation, $\varepsilon\rho^2(S)$ may be appropriate when ratings are used for the diagnosis of instructional problems. Here specific information is sought by the instructor as to where his/her instructional problems may lie. Generalizing over students is still important here, since ones interest in instructional problems would certainly transcend a particular class of students, especially since any corrections made will be experienced by subsequent classes. However, generalizing over items seems to be of less importance, due to the specific nature of the information sought. One could argue that even so, one would like to generalize to an infinite set of items within the particular restricted domain and thus would prefer use of $\varepsilon\rho^2(S,I)$. We feel that this would represent an overly conservative decision, however.

Finally, since not generalizing over students seems to be inappropriate for either of the above uses of instructional ratings, of what use is $\varepsilon\rho^2(I)$? For student course evaluations, its usefulness is very limited. Possibly in the instrument development stage, when one is pilot testing items and wants to be assured of a homogeneous and internally consistent set, $\varepsilon\rho^2(I)$ could help guide decisions as to which and how many items need to be used. However, one would be ill-advised to use $\varepsilon\rho^2(I)$ to evaluate the dependability of the instrument in terms of its intended application. Of course, within other applications, such as class-room testing, $\varepsilon\rho^2(I)$ can be very important.

In terms of relative magnitude, $\varepsilon\rho^2(S,I)$ is necessarily smaller than both $\varepsilon\rho^2(S)$ and $\varepsilon\rho^2(I)$. The estimates of the expected observed score variances are the same for all three coefficients. The estimate of the universe score variance for $\varepsilon\rho^2(S,I)$, involves only one term, $\sigma^2(c)$, while the estimates of the universe score variance for $\varepsilon\rho^2(S)$ and $\varepsilon\rho^2(I)$ each include an additional non-negative term, $\frac{1}{k}$, $\sigma^2(ci)$ and $\frac{1}{n}$, $\sigma^2(s,cs)$ respectively. Thus both $\varepsilon\rho^2(S)$ and $\varepsilon\rho^2(I)$ yield inflated estimates of $\varepsilon\rho^2(S,I)$ if used for that purpose.

Analysis of Student Instructional Ratings Data

Data deriving from use of the University of Washington Instructional Assessment System (IAS) (Gillmore, 1974) were analyzed according to the random effects model presented in Table 1. From all classes in which IAS forms were used during the academic year, 1974-75, two sets of forty-two classes each were chosen subject to certain restrictions. First, each class was required to have at least fourteen students who completed the forms. Secondly, each sample was drawn such that it contained fourteen classes from each of three general areas, Humanities, Social Sciences, and Natural Sciences.

The IAS contains five evaluational forms. However, eleven items are common to every form. Since the sample contained users of more than one form,

only data from these eleven items were used. The items are found in Table 2, and all use the response categories: Excellent, Very Good, Good, Fair, Poor, and Very Poor; coded 5, 4, 3, 2, 1, and 0 respectively.

Insert Table 2 about here

To assure a balanced design, a subset of thirteen students was randomly selected from each class containing more than thirteen students. Thus, for each of two samples, forty-two classes were rated on the same eleven items by thirteen students. The students were, of course, nested within classes and crossed with items. Data from the two samples were analyzed separately, so that the results could be compared for stability.

The results of the analyses of variance and the estimates of the variance components for the two samples are presented in Table 3. The results of the two samples are very similar, indicating that the sample sizes are large enough to produce stable estimates of the variance components.

Insert Table 3 about here

The most striking feature of these analyses is that the variance attributable to sampling students ($\sigma^2_{b,cs}$) is much larger than the variance due to sampling of items ($\sigma^2_{(i)}$ and $\sigma^2_{(ci)}$). The variance component of items is very small, showing that our decision to restrict our attention to applications in which all students respond to the same set of items will not have a strong impact on generalizability estimates.

These results are similar, especially in terms of relative magnitude, to those reported in Kane, Gillmore, and Crooks (in press), who applied the same method to results of the University of Illinois Physics Questionnaire and

Course Evaluation Questionnaire. Of particular interest is their analysis of the general items of the Physics Questionnaire (see Crooks (1973) for details of the development of this instrument). These items were similar to those analyzed in the present study and used the same response categories. Estimated variance components were reported as follows: $\hat{\sigma}^2_{(c)} = .12$, $\hat{\sigma}^2_{(s,cs)} = .54$, $\hat{\sigma}^2_{(i)} = .05$, $\hat{\sigma}^2_{(ci)} = .05$, and $\hat{\sigma}^2_{(si,csi,e)} = .40$. The reader can see that these values are very similar to those reported in Table 3, even though the items were somewhat different, the institutional setting was different, and the sample was comprised of all Physics instructors.

In Table 4 all three generalizability coefficients, $\varepsilon\rho^2(S,I)$, $\varepsilon\rho^2(S)$, and $\varepsilon\rho^2(I)$, are presented for a large variety of possible D-studies. These coefficients are based on the values of the variance components from Sample 1 only. The generalizability coefficients for Sample 2 are essentially the same as those for Sample 1. They were not viewed as sufficiently informative or instructive to be included in this report. The interested reader can comput specific values from Sample 2 by using formulas 1, 2, or 3.

Insert Table 4 about here

The first thing the reader may note from Table 4 is that the values of both $\varepsilon\rho^2(S)$ and $\varepsilon\rho^2(I)$ are larger than the corresponding value of $\varepsilon\rho^2(S,I)$ for equivalent numbers of items and subjects. This is necessarily true because all three coefficients have the same denominator, but $\varepsilon\rho^2(I)$ and $\varepsilon\rho^2(S)$ each have an additional positive term in the numerator. In a few cases in Table 4, $\varepsilon\rho^2(S,I)$ is listed as equal to $\varepsilon\rho^2(S)$ (e.g., one subject, 100 items) because of rounding.

Values of $\varepsilon\rho^2(S,I)$ are plotted in Figure 1 as a function of the number of items for one, five, twenty-five and one-hundred students. As mentioned earlier, these curves do not follow the Spearman-Brown formula. It is clear that the magnitude of this coefficient is influenced more by the number of students than by the number of items. This is a direct result of the fact that $\sigma^2(s,cs)$ is eight times larger than $\sigma^2(ci)$. Increasing the number of items has a large effect at first, but the increase quickly levels out and increasing the number of items beyond ten has little effect. Notice, however, that with few students, no number of items produces a satisfactory level for $\varepsilon\rho^2(S,I)$. Increasing the number of students naturally has a pronounced effect on $\varepsilon\rho^2(S)$, because it decreases the size of both of the sources of error. By contrast, increasing the number of items beyond four or five has relatively little effect.

Figure 2 contains the values of $\varepsilon\rho^2(S)$ as a function of number of students, for one item and for one-hundred items. Each of these curves follows the Spearman-Brown formula. Increasing the number of students naturally has a pronounced effect on $\varepsilon\rho^2(S)$, because it decreases both sources of error for this coefficient. Increasing the number of items has very little effect, both because of the small size of the ci variance component and because it is part of the universe score variance and hence enters into both the numerator and the denominator.

The values of $\varepsilon\rho^2(I)$ are plotted as a function of the number of items for one and one-hundred students in Figure 3. Each of these curves also follow the Spearman-Brown formula. Adding additional items does not have the great effect on $\varepsilon\rho^2(I)$ that adding students had on $\varepsilon\rho^2(S)$. This is another ramification of the large student variance relative to item variance. With $\hat{\sigma}(s,cs)$ part of the universe score, values of $\varepsilon\rho^2(I)$ are high even for one item, and hence do not have much room to increase. The number of students does have an effect on $\varepsilon\rho^2(I)$, but its effect is confined mainly to very small numbers of items.

If we wish to consider either $\varepsilon\rho^2(S)$ or $\varepsilon\rho^2(I)$ as estimates of $\varepsilon\rho^2(S,I)$, we can see from either Table 4 or comparison of the figures that the former does not generally overestimate $\varepsilon\rho^2(S,I)$ to a great extent. The largest differences between the two are found where the number of items is small. This makes intuitive sense in that it is relatively safer to generalize to a finite set of items than to an infinite set when the number of items in the sample is very small. As the number of items becomes larger, the difference between two conceptions ought to lessen.

Unlike $\varepsilon\rho^2(S)$, $\varepsilon\rho^2(I)$ overestimates $\varepsilon\rho^2(S,I)$ rather grossly in all cases except where the number of subjects is large. However, for that case, the closeness of the two is due to the operation of a ceiling effect, since both coefficients are bounded from above at 1.00.

Discussion

We began this paper by arguing that the use of generalizability theory provides some real advantages over the use of traditional reliability coefficients. In the application of generalizability theory presented above, two primary advantages seem to emerge.

First, with generalizability we were able to clearly define and interpret three coefficients and trace the behavior of each with changes in numbers of items and/or students. A traditional reliability analysis would probably lead to two of these coefficients; internal consistency, corresponding to $\varepsilon\rho^2(I)$, and stability, corresponding to $\varepsilon\rho^2(S)$. The most important coefficient from our point of view, $\varepsilon\rho^2(S,I)$, might well be missed.

Furthermore, one could use the Spearman-Brown formula to assess changes in internal consistency as a function of the number of items and changes in stability as a function of the number of students. However, to assess the effect of changing the number of items on stability or the number of students on

internal consistency would require new studies within traditional reliability. Furthermore, there would be no simple way to study a reliability coefficient corresponding to $\varepsilon\rho^2(S,I)$ as a function of students and items simultaneously.

The second advantage which can be cited results from the multifaceted interpretation of error variance which is the basis for generalizability theory. Similar to the analyses reported by Kane, et al. (in press), our analyses show that the variability of student responses is much greater than the variability within students. This clearly indicates that increasing the number of students will have a greater impact on dependability than increasing the number of items. In fact, item sampling seems to make a very minor contribution to the total observed variance. This is partly a result of our conceptualization of the D-study design which results in exclusion of the variance component from the main effect of items. However, it should be noted that this variance component is estimated to be quite small in our data and thus its inclusion would have only a very small effect on the results.

There are several decision-making implications implicit in the theory and analyses presented above. Each of the three coefficients is appropriate for use within some purposes but inappropriate for others. We have already argued for use of $\varepsilon\rho^2(S,I)$ if one is using the results of student evaluations to provide a general measure of teaching/course effectiveness. Using the data for decisions such as tenure, promotion, and merit pay would seem to indicate use of this index of dependability. $\varepsilon\rho^2(I)$ is an especially inappropriate index to use since we would not be interested in evaluating a teacher/course offering for one single group of students. Also, $\varepsilon\rho^2(I)$ seriously over-estimates $\varepsilon\rho^2(S,I)$ in all cases except where the number of students is large. Since the same items are likely to be used in evaluating instruction over several years, it is conceptually less objectionable to

generalize over students only, and the data reported here indicate that the value of $\varepsilon\rho^2(S)$ is not as seriously inflated when considered an estimate of $\varepsilon\rho^2(S,I)$.

Because the variability of student responses is much larger than the variability within students, it is more important to have a large sample of students than it is to have a large sample of items for dependability. Informal sampling of student opinion is likely to involve only a small number of students and for this reason alone, will tend to be unreliable. Also, informal sampling increases the possibility of the introduction of systematic bias in individual cases. For both of these reasons any decisions based on student opinion should be based on a sample of each class which is as exhaustive as possible. If student opinion is to be used as a measure of instructional effectiveness, a systematic procedure that insures a high rate of student response should be used. In order to obtain this exhaustive sample, it may be necessary to have the questionnaire administered on several occasions (several class sessions), where the administration in each session would be restricted to students who had not filled out the questionnaire in any previous session.

While the number of subjects is clearly more important than the number of items, our data indicate that the measure of teaching effectiveness should be based on about five items. Further increases in the number of items results in little increase in $\varepsilon\rho^2(S,I)$.

Student evaluative information is also used for diagnostic purposes. We have argued that for this purpose, one can be interested in the information yielded by specific items and not by items as representative of an infinite set. Since we still want to generalize over students, $\varepsilon\rho^2(S)$ seems to be an appropriate measure of dependability.

Using $\varepsilon\rho^2(S)$, our data indicate that one item will provide adequate dependability if the number of students is large. In fact, increasing the number of items beyond one offers little advantage. This implies that individual items can provide dependable diagnostic information and subscales of two or more items within the same specific domain are not necessary for providing dependable diagnostic information. As for students, the above comments concerning the advisability of obtaining ratings from an exhaustive sample within a class apply equally to this usage.

Considering only the domain from which the items of this study can be considered a sample, values of $\varepsilon\rho^2(I)$ indicate that item internal consistency is not a serious problem. Very small numbers of items combine to gove dependable results. For this reason as well as from the conceptual arguments made earlier, development of evaluation instruments should probably proceed from bases other than maximization of $\varepsilon\rho^2(I)$. For example, it is more important to construct or choose items within domains which relate to the precise information sought, than to choose items from a large pool on the basis of a factor analytic solution. It just does not seem important to have a number of items which load on the same "evaluative dimension". Relaxing the emphasis typically given internal consistency might result in some increase in $\sigma^2(ci)$, but this could be counteracted by increases in the number of items.

References

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of bahavioral measurements. New York: John Wiley and Sons, 1972.

Cronbach, L. J., Ikeda, M., & Avner, R. A. Intraclass correlation as an approximation to the coefficient of generalizability. Psychological Reports, 1964, 15, 727-736.

Crooks, T. J. Teaching and Learning in an introductory college Physics course. Unpublished doctoral dissertation, University of Illinois, 1973.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Gillmore, G. M. A brief description of the University of Washington Instructional Assessment System (EAC Report 276). Seattle: University of Washington, Educational Assessment Center, 1974.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Kane, M. T., and Brennan, R. L. The generalizability of class means. Review of Educational Research (in press).

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement (in press).

Table 1

Summary of Random Effects ANOVA for Design V-B

Source of Variance	Sources as confounded	df	E(MS)
c	c	j-1	$\sigma^2(e) + k\sigma^2(s,cs) + n\sigma^2(ci) + nk\sigma^2(c)$
Within c			
s			
cs	s,cs	j(n-1)	$\sigma^2(e) + k\sigma^2(s,cs)$
i	i	k-1	$\sigma^2(e) + n\sigma^2(ci) + jn\sigma^2(i)$
ci	ci	(j-1)(k-1)	$\sigma^2(e) + n\sigma^2(ci)$
Within ci			
si			
csi,e	si,sci,e	j(n-1)(k-1)	$\sigma^2(e)$

Table 2

Common Items of the Instructional Assessment System

The Course as a Whole Was:

The Course Content Was:

The Instructor's Contribution to the Course Was:

The Instructor's Effectiveness in Teaching the Subject Matter Was:

Use of Class Time Was:

Instructor's Interest in Whether Students Learned Was:

Amount you Learned in the Course Was:

Relevance and Usefulness of Course Content Is:

Evaluative and Grading Techniques (Tests, Papers, Projects, etc.) Were:

Reasonableness of Assigned Work Was:

Clarity of Student Responsibilities and Requirements Was:

Table 3

Analysis of Variance Summary Tables and Variance Components Estimates for Both Samples

	Source	DF	SS	MS	$\hat{\sigma}^2$
	c	41	1039.29	25.35	.13
	s,cs	546	2825.64	5.18	.43
Sample 1	i	10	74.65	7.46	.01
	ci	410	455.07	1.11	.05
	si,csi(e)	5460	2505.00	.46	.46
	c	41	1041.22	25.40	.12
	s,cs	546	3054.51	5.59	.46
Sample 2	i	10	179.85	17.99	.03
	ci	410	530.37	1.29	.06
	si,csi (e)	5460	2671.78	.49	.49

Table 4

Generalizability Coefficients ($\varepsilon\rho^2(S,I)$, $\varepsilon\rho^2(S)$, and $\varepsilon\rho^2(I)$ for Various D-Studies

Number of Students		Number of Items						
		1	2	5	10	25	50	100
1	$\varepsilon\rho^2(S,I)$	.12	.16	.20	.21	.22	.23	.23
	$\varepsilon\rho^2(S)$	.17	.19	.21	.22	.23	.23	.23
	$\varepsilon\rho^2(I)$	.52	.69	.85	.92	.96	.98	.99
2	$\varepsilon\rho^2(S,I)$	.21	.27	.32	.35	.36	.37	.37
	$\varepsilon\rho^2(S)$	.29	.32	.35	.36	.37	.37	.38
	$\varepsilon\rho^2(I)$	.55	.71	.86	.92	.97	.98	.99
5	$\varepsilon\rho^2(S,I)$	.36	.45	.53	.56	.59	.59	.60
	$\varepsilon\rho^2(S)$	.50	.54	.57	.59	.59	.60	.60
	$\varepsilon\rho^2(I)$	.60	.75	.88	.94	.97	.99	.99
10	$\varepsilon\rho^2(S,I)$	.48	.59	.68	.71	.74	.74	.75
	$\varepsilon\rho^2(S)$	.67	.70	.73	.74	.75	.75	.75
	$\varepsilon\rho^2(I)$	.64	.78	.90	.95	.98	.99	.99
25	$\varepsilon\rho^2(S,I)$	.60	.72	.81	.84	.87	.88	.88
	$\varepsilon\rho^2(S)$	.83	.85	.87	.88	.88	.88	.88
	$\varepsilon\rho^2(I)$	.68	.81	.91	.96	.98	.99	.99
50	$\varepsilon\rho^2(S,I)$	.66	.77	.86	.90	.92	.93	.93
	$\varepsilon\rho^2(S)$	.91	.92	.93	.93	.94	.94	.94
	$\varepsilon\rho^2(I)$	.70	.82	.92	.96	.98	.98	1.00
100	$\varepsilon\rho^2(S,I)$	.69	.80	.90	.93	.95	.96	.96
	$\varepsilon\rho^2(S)$	.95	.96	.96	.97	.97	.97	.97
	$\varepsilon\rho^2(I)$	.71	.83	.92	.96	.98	.99	1.00

Figure 1: $\epsilon\rho^2(S,I)$ as a Function of Number of Items for Various Levels of Class Size.

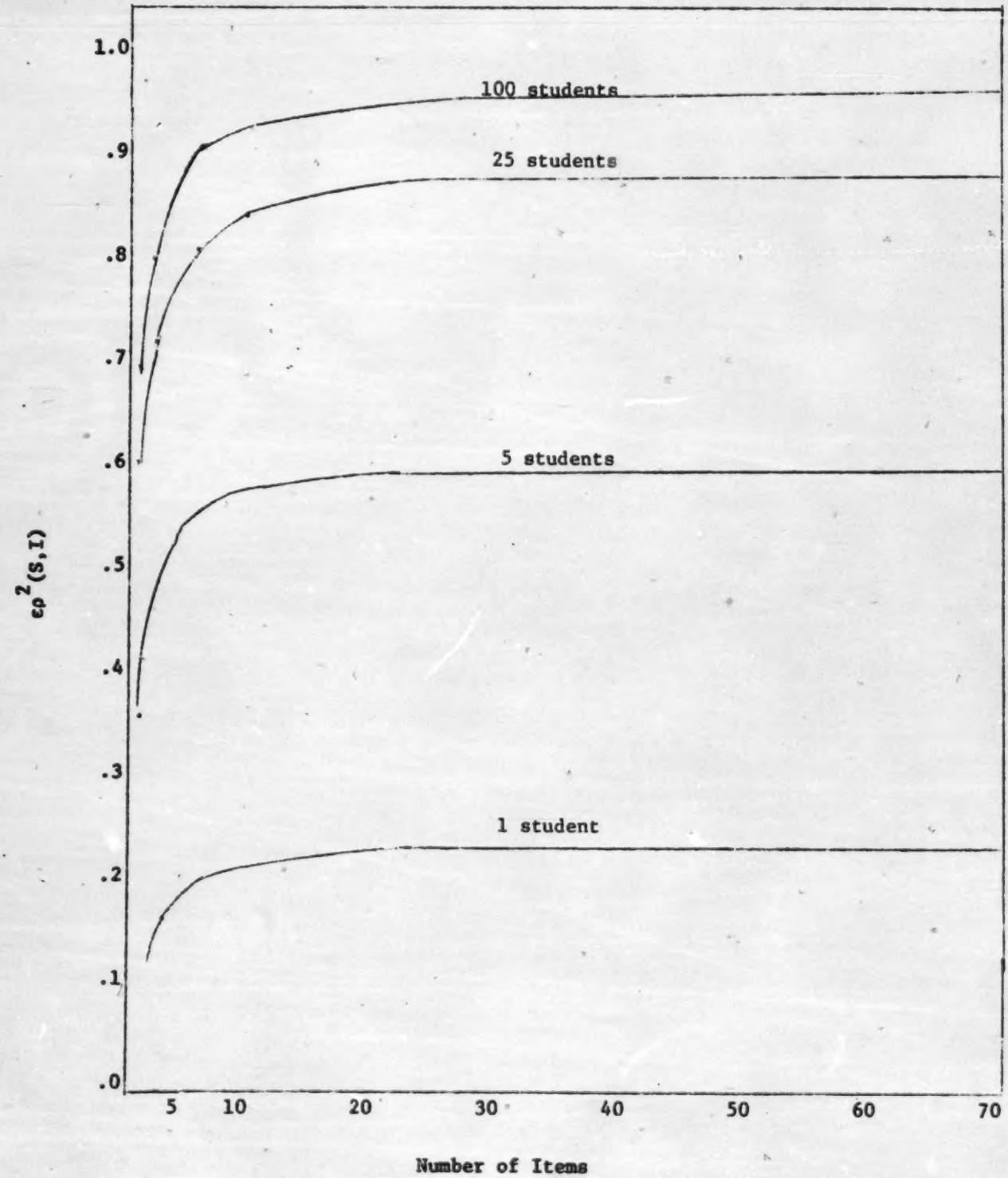

Figure 2. $\varepsilon\rho^2(S)$ as a Function of Class Size for Two Levels of Items.

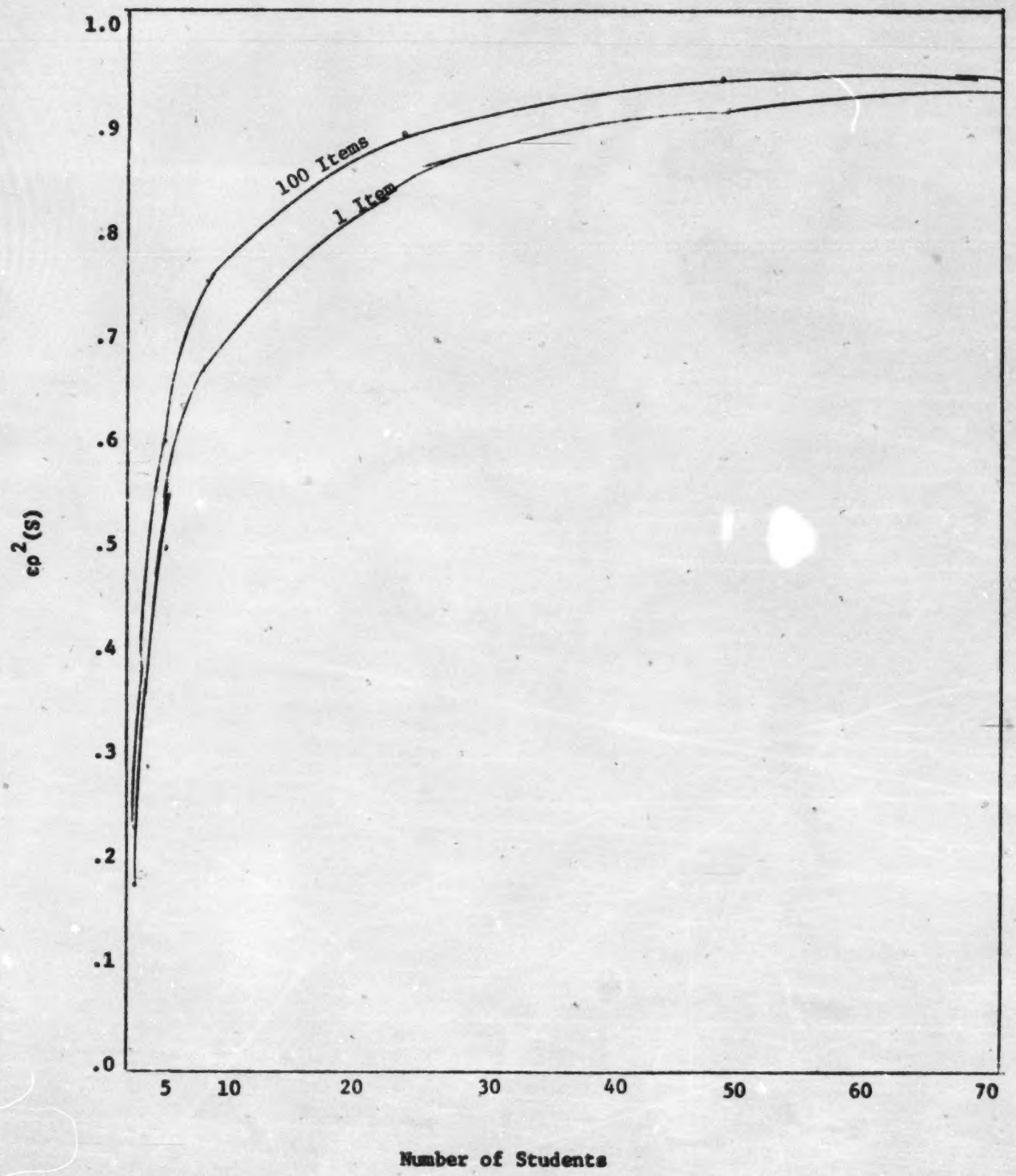

Figure 3: $\epsilon\rho^2(I)$ as a Function of Number of Items for Two Levels of Class Size.